The ethics of sample size: Two-sided testing and one-sided thinking

Knottnerus, J.A.; Bouter, L.M.

published in
Journal of Clinical Epidemiology
2001

DOI (link to publisher)
10.1016/S0895-4356(00)00276-6

document version
Publisher's PDF, also known as Version of record

ELSEVIER

Journal of Clinical Epidemiology 54 (2001) 109–110

Journal of Clinical Epidemiology

COMMENTARY

The ethics of sample size: Two-sided testing and one-sided thinking

J. André Knottnerus[a,*], Lex M. Bouter[b]

[a]*Netherlands School of Primary Care Research, University of Maastricht, P.O. Box 616, 6200 MD Maastricht, The Netherlands*
[b]*Institute for research in Extramural Medicine, Vrije Universiteit, Amsterdam, The Netherlands*

Received 2 March 2000; received in revised form 2 May 2000; accepted 5 May 2000

While the ethical debate on clinical trials is often focused on informed consent and related issues, estimation of the required sample size is usually not evaluated as to its ethical implications. However, these implications can be considerable in terms of subjects burdened with experimental interventions and of the surplus number of needed primary endpoints or adverse outcomes before superiority or inferiority of an intervention is accepted.

In a standard randomised controlled trial the outcome of a group subjected to the studied intervention of primary interest, henceforth to be designated as "principal" intervention, is compared with the outcome of a group subjected to a reference intervention. To collect sufficient evidence to detect a clinically relevant difference between both interventions, if present, a minimum number of subjects per group should be included. At the same time, from an ethical point of view and to warrant an efficient use of resources, it is also required that the sample size be not larger than needed.

The estimation of the minimum sample size requires the specification of the minimal difference in outcome (delta) that would be clinically important to be detected. In addition, investigators must specify an acceptable type I error (the probability of detecting a significant difference when the treatments are really equally effective), an acceptable type II error (the probability of not detecting a significant difference when there really is a difference of magnitude delta or larger, that is, 1 minus power), and an (evidence-based) estimation of the distribution of the outcome parameter in the reference group [1]. Finally, one should explicitly choose between two-sided or one-sided statistical testing. Unfortunately, the importance of the latter decision is often neglected [2–5].

In the two-sided option one is able to detect the predefined relevant difference, if present, in both directions. That is, if the principal treatment is an amount delta better than the reference, this will probably be found, but also a similar advantage of the reference over the principal intervention can be detected. The one-sided option would only allow a one-sided evaluation, (e.g., whether or not the principal is an amount delta better than the reference treatment), without testing whether the reference is superior to the principal intervention.

In the current practice of estimating or reviewing the needed sample size, two-sided testing is virtually always used, while one-sided testing is mostly considered unacceptable because it does not account for the possibility that the reference treatment might be better [6]. However, in order to detect a similar difference in outcome between both interventions, two-sidedness is paid for by a considerably larger sample size. For example, accepting a type I error of 0.05 and a type II error of 0.20, and assuming a cumulative incidence during follow-up of 20% of the primary endpoint in the reference group, the detection of a clinically relevant difference in cumulative incidence of 10% favouring the principal intervention needs a minimum sample size of 157 per group in case of one-sided testing; and of 199 per group in case of two-sided testing. So, the latter would imply an additional 84, that is $2 \times (199 - 157)$, patients to be included in the trial.

If the incidence of the primary endpoint in the reference group is assumed to be 50%, *ceteris paribus*, the corresponding numbers of study subjects per group to detect a similar reduction of 10% would be 305 and 388, respectively [7]. This means that the two-sided approach requires inclusion of an additional 166 subjects, with a notable difference in burdening of patients, trial feasibility and cost. Moreover, if a fatal outcome is studied, in the two-sided case of the latter example an average of 8 casualties (31 in the one-sided case versus 39 in the two-sided case) would be additionally needed to conclude that the principal treatment would be better than the reference.

In choosing between a one-sided or a two-sided approach it should be recognized that a research hypothesis is not random shooting but expresses scientific uncertainty regarding

* Corresponding author. Tel.: 0031 43 388 2319; fax: 0031 43 367 1458.

E-mail address: Andre.Knottnerus@HAG.Unimaas.NL (J.A.Knottnerus)

PII: S0895-4356(00)00276-6

a plausible, potentially clinically important effect [8–10]. Before a trial one is not totally ignorant of what is to be tested. The problem at issue is generally not: "is the principal treatment better than the reference, or does it equal the reference, or is it worse than the reference?". Rather, the question is "whether the principal treatment is indeed better than the reference," considering that an advantage of the principal over the reference is a reasonable but not sufficiently tested assumption. Thus, a research question is often hypothesis-driven and typically "one-sided."

If one cannot exclude the possibility that, unexpectedly, the principal treatment would be worse than the reference, it can still be important to test this possibility in a two-sided approach. However, this is only clinically meaningful if such a result would change current practice. If the expected balance between favourable effects and adverse effects of the principal treatment is such that only clear superiority in primary endpoints is relevant, not being able to demonstrate a relevant advantage of the principal treatment is evidence in favour of the reference.

Even if interventions have similar effects as to primary outcome, complications, and side effects, their burdens for the patient may be considerably different. For example, if carotid endarterectomy in asymptomatic patients is not shown to produce substantially better outcomes than the non-surgical approach, the latter will be preferred because it is less invasive for the patient [11]. And in the absence of a demonstrated advantage of coumarin over aspirin in preventing stroke in primary care patients with non-valvular atrial fibrillation, aspirin is the drug of choice since coumarin treatment requires frequent blood testing and makes the patient more dependent on health care [12].

Given these considerations, one-sided testing and a corresponding sample size estimation can be proposed as the preferred approach if: (1) the scientific hypothesis to be tested is obviously one-sided, or if (2) only a clear advantage in effect of the principal over the reference intervention would have consequences for practice, for example, if the principal intervention implies a more burdensome regimen for the patient [13]. Accordingly, if the reference intervention is refraining from treatment or an inactive, placebo therapy, a useful option is that an active principal intervention should be shown to be clearly better. Therefore, in trials with "non-treatment" or placebo reference groups, as are often carried out to test newly developed interventions, the one-sided approach would be adequate.

Of course, we must recognize that in deciding upon first-choice treatments in practice, things are more complex than making inferences based on just one trial. The body of knowledge generally consists of a mosaic of studies, also providing insight in the magnitude of effects in relevant subgroups. Supported by meta-analysis and statistical pooling, this may yield a powerful evidence base for clinical choices. In fact, this brings the issue of optimal sample size estimation on a supra-individual trial level: what degree of uncertainty, based on a systematic review of previous trials, is sufficient to justify an additional study on the same topic? Regarding the research hypothesis, prior evidence and the clinical implications of the treatment under study, should the sample size and analysis of a new study be based on a one-sided or a two-sided approach? How large should a new study be to achieve enough power in addition to the already performed studies? And how to prospectively design a set of collaborative studies of sufficient size to be successfully pooled afterwards?

In the meantime, however, for individual trials, especially into new interventions not previously studied, a one-sided view seems sensible.

References

[1] Pocock SJ. Clinical trials, a practical approach. Chichester: John Wiley & Sons, 1983.
[2] Dunnett CW, Gent M. An alternative to the use of two-sided tests in clinical trials. Stat Med 1996;15:1729–38.
[3] Peace KE. The alternative hypothesis: one-sided or two-sided? J Clin Epidemiol 1989;42(5):473–6.
[4] Bland JM, Altman DG. One and two sided tests of significance. BMJ 1994;309:248.
[5] Koch GG. One-sided and two-sided tests and p values. J Biopharm Stat 1991;1:161–70.
[6] Walker AM. Low power and striking results—a surprise but not a paradox. N Engl J Med 1995;332:1091–2.
[7] Machin D, Campbell MJ. Statistical tables for the design of clinical trials. Oxford: Blackwell Scientific Publications, 1987.
[8] Avins AL. Can unequal be more fair? Ethics, subject allocation, and randomised clinical trials. J Medical Ethics 1998;24:401–8.
[9] Feinstein AR, Concato J. The quest for "power": contradictory hypotheses and inflated sample sizes. J Clin Epidemiol 1998;51:537–45.
[10] Knottnerus JA. Gezondheidszorg in extramurale settings. Ethiek en Recht in de Gezondheidszorg 1997;IV:151–94.
[11] Benavente O, Moher D, Pham B. Carotid endarterectomy for asymptomatic carotid stenosis: a meta-analysis. BMJ 1998;317:1477–80.
[12] Hellemons BSP, Langenberg M, Lodder J, Vermeer F, Schouten HJA, Lemmens Th, Ree JW van, Knottnerus JA. Primary prevention of arterial thromboembolism in non-rheumatic atrial fibrillation in primary care: randomised controlled trial comparing two intensities of coumarin with aspirin. BMJ 1999;319:958–64.
[13] Riffenburgh RH. Statistics in medicine. San Diego, CA: Academic Press, 1999.